Owen Walls, December 19th, 2024

Outside the Chinese Room: How We Talk Past One Another

ABSTRACT:

This paper, “Outside the Chinese Room: How We Talk Past One Another,” examines the philosophical discourse surrounding John Searle’s Chinese Room Argument (CRA), with a particular focus on the miscommunications and conceptual divides that persist between its critics and its proponents. Searle’s thought experiment challenges the claims of “Strong AI” by asserting that syntax, the manipulation of formal symbols, cannot generate semantics or genuine understanding. Critics have responded with varying approaches, but many of these critiques fail to effectively engage with the intuitive underpinnings of Searle’s argument. Through an analysis of Georges Rey’s Systems Reply in “What’s Really Going on in Searle’s Chinese Room?” and Margaret Boden’s Robot Reply in “Escaping from the Chinese Room,” this paper diagnoses the debate as a clash between intuition-driven arguments and rational counterarguments. Searle’s reliance on intuition underscores his claim that understanding is absent in the Chinese Room, while rational critiques, which prioritize observable functional equivalence, often bypass the intuitive core of Searle’s argument. This paper argues that these rational responses “talk past” Searle by attempting to counter his intuitions with purely rational argumentation. By introducing the Speed Reply and analyzing large language models (LLMs) as a contemporary counterexample, this paper offers a framework for bridging this epistemological divide and advancing the discussion beyond its current impasse.

INTRODUCTION:

In his 1980 paper “Minds, Brains, and Programs,” John Searle argues that computer programs are not analogous to human cognition on the grounds that syntax, or formal symbol manipulation, and semantics, the attribution of meaning to those symbols, lack an intrinsic relationship that would allow the latter to emerge from the former. He claims that understanding arises from something unique to the brain as a biological organism, something no computer can approximate. This view, known as biological naturalism, challenges the functionalist doctrine that mental states are defined by their functional roles rather than the physical substrate in which they are realized. Functionalism, which underpins the claims of “Strong AI,” holds that cognition can be realized in any system that performs the appropriate functions, regardless of its material composition. For functionalism, what matters is not what a system is made of but what it does.

Searle’s critique of functionalism is rooted in his novel and highly contested Chinese Room thought experiment. In the scenario, Searle imagines himself, an English monoglot, locked in a room manipulating Chinese symbols using a rulebook. Though the responses appear meaningful to an outside observer, Searle himself understands nothing of the language. This disjunction, according to Searle, reveals that functional equivalence, the ability to produce outputs indistinguishable from those of a human mind, does not entail genuine understanding. Instead, Searle claims, understanding is contingent on the unique biological processes of the brain. Searle’s Chinese Room has elicited a range of responses, most notably the Systems and Robot Replies, which aim to defend functionalism against his critique. These responses argue that understanding must arise from the system as a whole, including the rulebook, the room, and Searle himself, or from embodiment, where a system demonstrates understanding by interacting with the world through sensors and motors. However, these critiques fail to effectively engage Searle’s argument. Searle’s critique is rooted in intuition, a gut-level assessment of what understanding “feels like,” whereas his critics respond with rational,

behaviorist arguments that focus on measurable inputs and outputs. These incommensurable modes of thought leave the debate at an impasse.

This paper argues that diagnosing this epistemological divide, intuitional versus rational thinking, is key to understanding why the debate remains unresolved. By clarifying how Searle's critics talk past him, I propose a method for engaging the implications of his argument while still granting his intuitions. Specifically, I argue that by introducing time as a functional variable, we can demonstrate that the Chinese Room and its counterpart, the English Room, are not truly functionally equivalent. The time required for Searle to manipulate symbols in the Chinese Room introduces a measurable difference that undermines the claim of functional equivalence. This diagnosis weakens the anti-functionalist implications of Searle's argument without attempting to refute his intuitions about the Chinese Room itself. Finally, I will consider the counterexample of large language models (LLMs), whose rapid and convincing outputs appear to challenge the Speed Reply. By analyzing the relationship between intuition and functional equivalence, I argue that LLMs ultimately reinforce the diagnostic framework and highlight the necessity of engaging commensurable modes of thought.

THE CHINESE ROOM:

The focal point of our discussion, Searle's Chinese Room, is constructed as follows: Searle asks us to imagine that he, an English speaker, is locked in a room with only a book containing a program to associate Chinese inputs with their appropriate Chinese outputs. The only medium between him and the outside world is a slot where he can receive Chinese scripts and send Chinese scripts out. As an English monoglot, the scripts are intelligible to him only syntactically as "meaningless squiggles";[1] in other words, he can identify them only by their shape. The Chinese symbol for Egg Foo Young, 芙蓉蛋, would not

[1] John Searle, "Biological Naturalism," in The Blackwell Companion to Consciousness, ed. Susan Schneider and Max Velmans (Malden, MA: Wiley-Blackwell, 2017), https://doi.org/10.1002/9781119132363.ch23,1.

refer to a dish of egg and vegetable but to three collections of horizontal and vertical lines distinct only in their length, number, and orientation.

Searle then proposes two situations. In the first, he is handed questions written in Chinese and tasked with answering appropriately, also in Chinese; in the second, he is asked questions in English, to be responded to in English. In the Chinese situation, he would compare the shapes of the symbols with the shapes of the symbols present in his program and determine which symbols he should output as appropriate responses to the input that preceded it. The details of Searle's program are nebulous. At its weakest, it could be imagined as a simple and unbelievably long lookup table that merely correlates every possible input/output, question/answer pair. But at its most robust, it could be a program consisting of numerous subprograms, like systems of quantifiable modal logic and preference orders, that more closely approximate a human mind.[2] As long as they are syntactic, Searle would argue that it doesn't matter what rules the program instantiates; the key is that he never accesses any of the semantic meaning the symbols might hold for a Chinese speaker who truly understands them. In Searle's view, the project of understanding would be no more feasible than learning Chinese from a Chinese/Chinese dictionary. In the second situation, he is handed questions in English, which he understands completely and answers appropriately, as he would any question in his native language. The efficacy of Searle's thought experiment relies on the dichotomy between the Chinese and English scenarios. In both cases, from the outside of the room, it would appear as though the questions were understood. Despite the room in each situation being functionally indistinguishable, in one case, Searle understands the question, and in the other, as far as we can intuit, he does not.

---

[2] Georges Rey, "What's Really Going on in Searle's 'Chinese Room,'" Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 50, no. 2 (September 1986): 169–185, https://www.jstor.org/stable/4319853.

FUNCTIONALISM:

This disjunction between the functional equivalence of the rooms and the presence of genuine understanding, according to Searle, highlights a critical flaw in functionalism, which is the idea that mental states are defined by their functional roles, how they interact with inputs, outputs, and other mental states, rather than by the physical substrate in which they are realized. This principle of multiple realizability underpins functionalism's explanatory power, allowing for the attribution of mental states to a wide range of systems, from human brains to AI programs, as long as they perform the necessary cognitive functions. Functionalists argue that "being is doing;"[3] if a system can do as a mind does, it can be as a mind is, regardless of what constitutes it.

Searle's Chinese Room directly challenges this view. For Searle, producing outputs that are indistinguishable from those of a native speaker, functional equivalence, is not sufficient to attribute understanding. In the Chinese Room, the individual following the rulebook manipulates symbols without ever accessing their meaning, demonstrating that syntax alone cannot produce semantics. From Searle's perspective, something more is needed to bridge the gap between symbol manipulation and meaningful understanding. Searle introduces biological naturalism to fill this gap. He argues that understanding is not merely a matter of computational processes but is rooted in the brain's unique biochemical makeup. For Searle, the intentionality that underpins understanding is analogous to digestion or photosynthesis: it is an emergent biological phenomenon entirely dependent on the specific physical processes of the brain. Unlike functionalism, which allows for mental states to be realized in any system that performs the appropriate functions, biological naturalism holds that certain phenomena, like genuine understanding, can only be realized in systems with the right biological constitution.[4]

---

[3] Thomas W. Polger, "Functionalism," *Internet Encyclopedia of Philosophy*, ed. James Fieser and Bradley Dowden, University of Cincinnati, accessed December 19, 2024, https://iep.utm.edu/functionalism/.

[4] Searle, "The Chinese Room," 6.

INTUITIONAL VS RATIONAL THINKING:

As we diagnose this debate, the epistemological distinction between intuitional and rational thinking is central, as they cause the philosophers who engage in these discussions to explore different and incompatible parts of the picture. Intuition provides direct access to the inside of the rooms, whereas rationality allows philosophers to make inferences based on what we know happens outside the rooms. Intuitional thinking is characterized by immediate, subjective judgments that arise from an instinctive sense of plausibility. Relying on gut-level assessments of what feels correct or intuitively evident, intuitions are resistant to rational justification. Searle's engagement with his Chinese Room is firmly grounded in intuition. He asks us to imagine ourselves in the room and recognize that, even though the system produces outputs indistinguishable from those of a native speaker, if you were standing in the room, you wouldn't understand Chinese. This disjunction between behavior and subjective experience is revealed only through our sense of intuition, we do not need formal reasoning to grasp the lack of understanding within the room; we simply feel it.

Rational thinking, by contrast, relies on formal logic, explicit reasoning, and measurable evidence to justify claims. It is the dominant mode of thought in critiques like the Systems and Robot Replies, which focus on demonstrating that understanding must arise from what we know based on observable behavior from outside the room. Rational thinking lends itself to arguments based on observable data or deductive reasoning, but it struggles to address the subjective, immediate insights that intuitional thinking provides. The rational critiques of the Chinese room make logical assertions about the room, ignoring that the gravity of Searle's argument lies in the disjunction between what is logically assertable and what we can intuit is happening.

SYSTEMS REPLY:

The purpose of the Chinese room, to demonstrate the necessity of biological naturalism, is forgotten by Searle's critics, who fruitlessly attempt to rationalize his intuitions. Searle's argument that there is a gap that needs to be closed by his biological naturalism is fundamentally as follows: In both the Chinese and English cases, understanding appears externally to occur, but we can use our intuition to determine that in one case, it really does occur, and in the other, it really does not. The Systems Reply attempts to disrupt this relationship by claiming that, despite our intuition, understanding actually occurs in both cases. They contend that in his example, in-the-room Searle does not represent the whole of the native Chinese speaker he's emulating but is merely their "central processing unit."[5] The entire system, constituted by Searle, his book, and whatever else exists in his imagined room, whether it be "scratch paper and pencils" or "data banks"[6] of Chinese symbols, must, as a whole, understand Chinese. Rey adopts a version of this reply in his What's Really Going on in the Chinese Room as he attempts to defend the thesis of Strong AI against Searle's criticism. Ironically, he charges Searle with almost the same thing I will charge him with here. Rey argues that Searle's critique is ineffectual because it strawmans the thesis of Strong AI, misattributing a behaviorist tilt. He claims that Searle regards the thesis of Strong AI as in accordance with that of the Turing Test; namely, that teletype input/output equivalence with a human being is a sufficient condition for sincere understanding. Remember, a sufficient condition is alone enough to bring about the consequent, in this case, implying that purely observable behavior necessitates internal understanding. This, however, is not the claim of Strong AI, which is chiefly concerned with whether the right sort of things happen inside the room. Importantly, though, the force of Searle's argument has very little to do with the Turing Test. Yes, both rooms pass the test, but only one understands. We know this because we know what happens inside the room. Searle's

[5] Rey, "What's Really Going on in Searle's 'Chinese Room,'" 173.
[6] John Searle, "The Chinese Room," excerpted from Minds, Brains, and Programs (1980), in The Philosophy of Artificial Intelligence, ed. Margaret A. Boden (Oxford: Oxford University Press, 1990), 3.

example aptly cuts right through the behavioral test, but Rey's critique does not. In his attempt to address Searle's intuitions with behaviorist argumentation, he falls into the same trap he accuses Searle of.

Briefly, Rey amends Searle's program in the following ways: first, he suggests that the book contains "full recursive grammatical and semantical rules." The latter, he seems to believe, could come from the former, although exactly how this happens, or doesn't, is painfully unargued. Searle would likely grant that the book should contain any and all required syntactic rules, but it is unclear how semantical rules would fit inside an intra-linguistic framework. Imagine trying to ascertain the meaning of any one word from a Chinese/Chinese dictionary. A central tenet of Searle's thought experiment is that syntax and semantics are non-fungible; no set of rules could allow in-the-room Searle to glean any semantic meaning at all from the shapes of the symbols he takes as inputs. However, Rey asserts that these modified rules would put in-the-room Searle in a position to determine semantic meaning, a position that he suggests could be realized with his second modification.

Attempting to concede Searle's biological requisite, he asks that in-the-room Searle internalize the entire system, memorizing the rule book so that there is nothing in the system that is not also inside in-the-room Searle. It's important to note that the term 'internalization' might bias you toward attributing understanding where it isn't warranted. Here, the word is not meant to imply any sort of integration between the rules and the value systems or beliefs of in-the-room Searle; it merely means that the book and all other features of the room now exist entirely inside Searle in a physical sense. Even in response to Rey's now flesh-and-blood in-the-room Searle, with the ability to adequately answer questions in Chinese using only his brain, Searle would say that it is intuitively evident that as he recalls which symbol correlates with which symbol out, there is no semantic understanding occurring. Rather than approach the question intuitively, Rey asks literally, "What more could Searle conceivably want?" arguing rationally that everything that could be accounted for is accounted for; there is no gap. However,

Searle's very point is that there is something more happening in the English case than in the Chinese case. What exactly that is, Searle would say, is a product of the biochemical constitution of his brain entering the picture in the same way it does for human minds.

ROBOT REPLY:

The Robot Reply, which Margaret Boden champions in her "Escaping from the Chinese room," attempts to engage with Searles Chinese Room by emphasizing the role of embodiment. Boden argues that understanding does not arise merely from formal symbol manipulation but from embedding a computational system in a physical, interactive environment. By granting a system perceptual and motor capacities, such as cameras to "see" and robotic limbs to interact with objects, she argues that it can form causal relationships with the world. Boden contends that these causal relationships allow the system to meaningfully connect symbols to the external environment, thus bridging the gap between syntax and semantics. Her account is unconvincing, however because its efficacy rests solely on her assertion that "if the input-output behavior of such a robot were identical with that of human beings, then it would demonstrably understand."[7] Boden frames understanding in terms of observable behavior, what the system does from the outside. If a robot were to consistently act in ways indistinguishable from a human, recognizing objects, navigating its environment, and responding appropriately, Boden argues that this would demonstrate that the robot possesses genuine understanding. The problem, evidently, lies in the term "demonstrably." Boden assumes that demonstrating human-equivalent behavior is sufficient to attribute understanding, but this perspective fails to address the core of Searle's argument.

Searle's Chinese Room is not concerned with what is observable from the outside; rather, it relies on intuition to assess what mediates the input and output. Boden's focus on demonstration overlooks this key distinction. For Searle, the question is not whether the robot

[7] Margaret A. Boden, "Escaping from the Chinese Room," in The Philosophy of Artificial Intelligence, ed. Margaret A. Boden (Oxford: Oxford University Press, 1990), 257.

behaves as if it understands but whether there is any semantic understanding occurring within the system. To illustrate this, Searle returns to the thought experiment: imagine he is placed inside the robot's "brain," manipulating symbols according to rules that now account for sensory inputs and motor outputs. Despite the robot's apparently human-like behavior, Searle insists that he himself would remain oblivious to the meanings of the symbols he processes. Sensor input would simply be reduced to syntactic nonsense, and the appropriate output, even if it allowed the robot to pass the beansprouts and wok down the table upon request would still, to him, be meaningless. For Searle, no matter how convincingly the robot interacts with the world, its apparent intentionality is no more genuine than that of the original Chinese Room.

Boden, by looking only at external behavior, fails to address Searles intuitive argument that syntax cannot produce semantics. The issue is not whether the robot can convincingly simulate human understanding, it's whether anything beyond simulation is occurring internally. Boden's reliance on demonstrating understanding through behavior aligns her critique with rational, behaviorist frameworks, which prioritize observable phenomena over subjective insights. While her approach seeks to rationalize the CRA's conclusions, it inherently excludes the intuitional disjunction Searle identifies. For Searle, understanding involves something irreducible to observable behavior, a feature that intuition reveals but rational demonstration cannot capture. Boden's Robot Reply, like the Systems Reply, fails to engage with the core of Searle's argument.

SPEED REPLY:

Critics of my diagnoses might object that if my conclusions were generalized, they leave little room to engage with intuitional arguments at all. If intuitions are subjective experiences that we cannot engage rationally how is anyone to contend with claims like Searles? Is his argument inherently infallible? The answer, as I will attempt to demonstrate, is no. While Searle's thought experiment is constructed of nearly solely intuitional content, its relationship to

the broader philosophical claims of anti-functionalism and biological naturalism necessarily invoke rational operations. Thus, while we cannot rationally refute Searle's intuitions, we can still engage with the implications of his argument.

The disjunction between appearance and reality in the case of the Chinese and English rooms forms the foundation of Searle's argument against functionalism. He claims that since functionalism treats systems with the same input-output behavior as equivalent, it cannot account for the apparent difference between the English Room, which genuinely understands, and the Chinese Room, which does not. Therefore we need a different explanation, enter his biological naturalism. The Speed Reply challenges the functional equivalence of the rooms by introducing time as a variable. It argues that the processes required for the Chinese Room to function are inherently slower than those of the English Room. This difference in response time creates a measurable functional distinction, undermining Searle's claim that the two rooms are functionally equivalent.

Consider how the Chinese Room would operate in practice. Imagine the Chinese speaker outside the room asks, "Do you prefer noodles or egg foo young?" Inside the room, Searle must consult his rulebook to match the input symbols to their corresponding output, presumably following a lengthy series of instructions to determine which output symbols represent a preference. This process could take hours, or even days, before Searle produces the response, "egg foo young." To the Chinese speaker, the delayed response would be a curiosity; they might reasonably question what is happening inside the room, whether it truly understands. By contrast, in the English Room, where Searle does understand the question, the response would be nearly instantaneous. This difference in speed is not a trivial detail; it reflects a fundamental functional distinction between the two rooms. The English Room functions efficiently because it processes meaning directly, while the Chinese Room relies on cumbersome symbol manipulation. The two systems, therefore, are not functionally equivalent.

This distinction dissolves the puzzle Searle attempts to construct. Once we incorporate time, functionalism can adequately account for the observed differences between the two rooms. The English Room's ability to produce quick, convincing responses demonstrates genuine understanding within a functionalist framework. The Chinese Room's slow and unconvincing behavior, on the other hand, does not. The Speed Reply shows that, when time is considered, there is no need to invoke biological naturalism to explain understanding. Functionalism remains sufficient. This method respects Searle's intuition that the Chinese Room lacks understanding but still engages with his argument rationally by showing that the claim of functional equivalence between the two rooms is false. This approach bridges the epistemological divide between intuition and rationality, demonstrating that critics can meaningfully engage with Searle's anti-functionalist implications without dismissing the intuitions behind his thought experiment. By showing that functionalism can account for understanding, the Speed Reply removes the necessity of Searls would-be solution.

LLM REPLY:

Critics might wonder how contemporary instantiations of artificial intelligence, like large language models (LLMS), might enter this picture. If a model like ChatGPT produces convincing and near-instantaneous answers to questions, is it not plausible that a syntactic room could do the same? If the Speed Reply depends on the time required for the Chinese Room's symbol manipulation to highlight a functional difference, what are we to make of LLMs, which produce outputs quickly and fluently? Are LLMs functionally equivalent to the English Room? The answer is yes, they are, but that doesn't work against the Speed Reply. Despite their functional equivalence, the relevance of LLMs in this discussion is crucially limited by the fact that we cannot intuitively access their internal processes the same way we can for the Chinese Room.

In Searle's thought experiment, the Chinese Room is designed to contrast with the English Room. Our intuitions can identify that the Chinese Room lacks understanding because

we are privy to its internal processes, the symbol manipulation that produces its outputs. This transparency is what allows Searle to assert that the Chinese Room's understanding is merely an illusion. In the case of LLMs, however, we lack the same intuitional access. The LLM's processes are complex and opaque, rooted in vast statistical patterns that render them "black boxes." Since we cannot intuitively observe what happens inside the LLM Room, we cannot confidently assert that it does or does not understand based on intuition alone. This causes the debate to shift depending on which comparison is made. If we compare the English Room to the Chinese Room, functionalism remains explanatorily valid because the two systems are not functionally equivalent, as the Speed Reply demonstrates. The English Room responds quickly, reflecting genuine understanding, while the Chinese Room responds slowly, highlighting its lack of understanding. If, however, we compare the English Room to the LLM Room, the grounds for intuition disappear. Here, the two are functionally equivalent in speed and behavior, but because the LLM Room is a black box, we cannot intuitively assess whether it genuinely understands or merely simulates understanding. The lack of intuitional access means that any claims about the LLM's understanding must be made on different grounds. Thus, LLMs reinforce the diagnostic framework introduced by the Speed Reply. Either the dichotomy is:

1. English Room vs. Chinese Room: Here, intuition reveals that the two are not functionally equivalent, and functionalism accounts for understanding without invoking biological naturalism.
2. English Room vs. LLM Room: In this case, functional equivalence holds, but the lack of intuitional access shifts the debate to rational grounds, where intuition can no longer arbitrate questions of understanding.

In either case, the Speed Reply remains consistent: functional equivalence and intuitional access must align for the implications of the Chinese Room to hold. The consideration of large language models underscores the importance of respecting and engaging with intuitional and rational thinking independently.

CONCLUSION:

In this paper, I have argued that the persistent impasse in the debate around Searle's Chinese room arises from the incommensurability of the intuitional and rational modes of thought that underpin Searle's argument and his critics' responses. Searle's intuitions about understanding hinge on direct access to the inner workings of the Chinese Room, while rational critiques, such as the Systems and Robot Replies, evaluate the argument through functional equivalence and observable behavior. These two epistemological approaches consistently talk past each other, leaving the debate unresolved. After diagnosing this divide, the Speed Reply demonstrates how rational engagement with the implications of Searle's argument can bridge the gap. It accepts the validity of Searle's intuitions but shifts the focus to a rational critique of the functional equivalence between the Chinese and English Rooms. By introducing time as a functional variable, the Speed Reply reveals that the Chinese Room is not truly equivalent to the English Room. This undermines the need for Searle's turn to biological naturalism by preserving the coherence of functionalism. In clarifying the epistemological divide and demonstrating how rational engagement can proceed, this paper attempts to provide an alternative perspective with which to engage Searle's Chinese room.

BIBLIOGRAPHY:

1. Searle, John. “The Chinese Room.” Excerpted from Minds, Brains, and Programs (1980). In The Philosophy of Artificial Intelligence, edited by Margaret A. Boden, 1-7. Oxford: Oxford University Press, 1990.
2. Rey, Georges. “What’s Really Going on in Searle’s ‘Chinese Room.’” Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 50, no. 2 (September 1986): 169–185. https://www.jstor.org/stable/4319853.
3. Boden, Margaret A. “Escaping from the Chinese Room.” In The Philosophy of Artificial Intelligence, edited by Margaret A. Boden, 253-266. Oxford: Oxford University Press, 1990.
4. Larry Hauser. “Chinese Room Argument.” Internet Encyclopedia of Philosophy. Edited by James Fieser and Bradley Dowden. Accessed December 19, 2024. https://iep.utm.edu/chinese-room-argument/.
5. “The Chinese Room Argument.” Stanford Encyclopedia of Philosophy (SEP). First published March 19, 2004; substantive revision October 23, 2024. Edited by Edward N. Zalta. Accessed December 19, 2024. https://plato.stanford.edu/entries/chinese-room/.
6. Pinker, Steven. The Language Instinct: How the Mind Creates Language. New York: HarperCollins, 1994.
7. Churchland, Paul M., and Patricia Smith Churchland. “Could a Machine Think?” Scientific American 262, no. 1 (January 1990): 32–39. https://www.jstor.org/stable/24996642.
8. Gardner, Howard. The Mind’s New Science: A History of the Cognitive Revolution. New York: Basic Books, 1985.

9. Hocutt, Max. “Behaviorism.” Philosophical Intermediate Article. University of Alabama, Tuscaloosa, Alabama, USA. Accessed December 19, 2024. https://www.researchgate.net/publication/227992586.
10. Polger, Thomas W. “Functionalism.” Internet Encyclopedia of Philosophy. Edited by James Fieser and Bradley Dowden. Accessed December 19, 2024. https://iep.utm.edu/functionalism/.